Efficient Video Coding with Hybrid Spatial and Fine-Grain SNR Scalabilities

Rong Yan*[1], Feng Wu[2], Shipeng Li[2], Ran Tao[1] and Yue Wang[1]

[1] Dept of Electronic Engineering, Beijing Institute of Technology, Beijing 100081, China
[2] Microsoft Research Asia, #49 Zhichun Road, Haidian, Beijing, 100080, China

ABSTRACT

A flexible and effective macroblock-based framework for hybrid spatial and fine-grain SNR scalable video coding is proposed in this paper. In the proposed framework, the base layer is of low resolution and is generally encoded at low bit rates with traditional prediction based coding schemes. Two enhancement layers, i.e., the low-resolution enhancement layer and the high-resolution enhancement layer, are generated to improve the video quality of the low-resolution base layer and evolve smoothly from low resolution to high resolution video with increasingly better quality, respectively. Since bit plane coding and drifting control techniques are applied to the two enhancement layers, each enhancement bitstream is fine-grain scalable and can be arbitrarily truncated to fit in the available channel bandwidth. In order to improve the coding efficiency and reduce the drifting errors at the high-resolution enhancement layer, five macroblock coding modes with different forms of motion compensation and reconstruction, are proposed in this paper. Furthermore, a mode decision algorithm is developed to select the appropriate coding mode for each macroblock at the high-resolution enhancement layer. Compared with the traditional spatial scalable coding scheme, the proposed framework not only provides the spatial scalability but also provides the fine granularity quality scalability at the same resolution.

Keywords: layered coding, scalable coding, spatial scalable coding, SNR scalable coding, macroblock coding mode, fine granularity scalable (FGS), progressive fine granularity scalable (PFGS)

1. INTRODUCTION

With the rapid development in computers and networks, more and more users expect to enjoy multimedia services through various PC or non-PC devices over the Internet or wireless networks. However, this kind of ubiquitous multimedia services post great challenges to the conventional video coding techniques. Since the current Internet is a heterogeneous network, the connection speed between servers and clients may vary in a wide range. This requires video coding techniques provide different video bit rates according to the available channel bandwidth. In general, many non-PC devices only have low-resolution screen and limited computational power. This further requires that video coding techniques provide efficient video representation with different resolutions and different decoding complexity. A straightforward solution to meet the above requirements would be to compress the same video sequence into many bitstreams for every possible bit rate, resolution and device complexity. The video server would choose an appropriate bitstream to transmit to an individual user according to the actual connection speed and device capability. Obviously, this would be a great waste of system resource. On the other hand, even if we could manage to store all these bitstreams in the server, in a dynamically changing non-QoS guaranteed network, we still could not provide the best available video quality with a single bitstream. Therefore, an active research topic in video coding field is how to efficiently compress video sequence with different scalabilities, such as rate, quality, temporal, spatial and complexity scalabilities.

Spatial scalable coding can provide video at different resolutions to fit in a wide range of applications. Some related techniques have been developed in the past years [1-7]. Recent video coding standards, such as H.263++ and MPEG-4, have also adopted spatial scalable coding techniques [8-9]. Fig. 1 illustrates a typical architecture of the traditional spatial scalable coding scheme adopted in MPEG-4 standard. Generally speaking, in a traditional spatial scalable coding scheme, there are two layers: low-resolution base layer and high-resolution enhancement layer. At the base layer, the

* This work has been done while the author is with Microsoft Research Asia.

Visual Communications and Image Processing 2002,
C.-C. Jay Kuo, Editor, Proceedings of SPIE Vol. 4671 (2002) © 2002 SPIE · 0277-786X/02/$15.00

first frame is an intra (I) frame, and other frames in a Group of Pictures (GOP) are forward prediction (P) frames, which are always predicted from the previous low-resolution I frame or P frame. At the enhancement layer, only the first frame is a forward prediction (P) frame predicted from the current up-sampled base layer, and other frames are bi-directional prediction (B) frames predicted from both the previous P frame (or B frame) at the enhancement layer and the up-sampled current frame at the base layer.

In the traditional spatial scalable coding, the base layer bitstream can be decoded independently to obtain the low-resolution video. When the available channel bandwidth is more than a certain bit rate, called as Switch Bit Rate (SBR) in this paper, users can receive the whole high-resolution enhancement layer bitstream and obtain the high-resolution video. However, either at the low-resolution layer or at the high-resolution layer, once encoded, the video quality is fixed even if there may be more channel bandwidth available for that layer. Obviously, the traditional spatial scalable coding only provides coarse scalability, which cannot flexibly and precisely adapt to channel bandwidth fluctuations. Moreover, due to the non-scalable (SNR) limitation in the encoded bitstreams (either the base layer or the enhancement layer), it is difficult to provide better video quality to those users with higher speed connection and more computational power. For instance, the client with low-resolution screen can only get the low-resolution base layer video with the pre-encoded quality, even though there may be more channel bandwidth available. For convenience, whenever it is possible, we denote the term *high-resolution* as *HR* and *low-resolution* as *LR* throughout this paper.

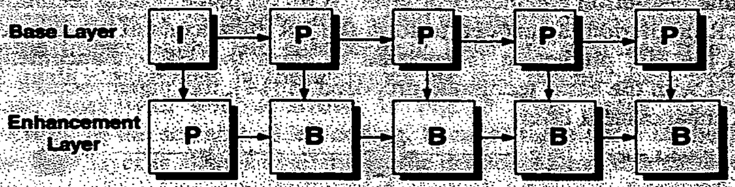


Fig. 1: The architecture of a traditional spatial scalable coding in MPEG-4.

Another scalable video coding technique known as MPEG-4 FGS (Fine Granularity Scalable) has become prominent in recent years due to its fine granularity SNR scalability [10]. Similar to the spatial scalable video coding, MPEG-4 FGS also compresses a video sequence into two layers of bitstreams. The base layer bitstream is generated by using prediction-based DCT transform technique at low bit rate. The residue between the original image and the reconstructed image of the base layer is compressed with bit plane coding technique to form the enhancement bitstream [11]. Since the bit plane coding technique provides an embedded bitstream and fine granularity scalability, the enhancement bitstream can be arbitrarily truncated to fit in the available channel bandwidth. The MPEG-4 FGS has only one motion compensation loop. The reference is always reconstructed from the low quality base layer to avoid possible drifting errors caused by truncation of the enhancement bitstream. Therefore, the coding efficiency of the enhancement layer is very poor due to the low quality motion prediction. Compared with the non-scalable video coding, the MPEG-4 FGS may lose 2.0dB or more in PSNR at the same bit rate. If the encoded video at the base layer were of low resolution, the coding efficiency of the enhancement layer would further decrease. On the other hand, as a non-trivial improvement to the MPEG-4, PFGS (Progressive Fine Granularity Scalable) video coding technique [12], while maintaining the fine granularity SNR scalability, significantly improves the coding efficiency of scalable bitstreams by using a second motion compensation loop with high quality references. Moreover, macroblock-based PFGS scheme [14] was proposed to further improve the coding efficiency by allowing different INTER coding modes for different macroblocks.

In fact, the problem of how to efficiently serve video data to the best of the capacity of different devices with different screen resolutions, different computational power and different connection speeds, cannot be solved by either spatial or SNR scalability alone. To deal with this problem, this paper proposes a novel framework with hybrid spatial and fine-grain SNR scalabilities. The proposed framework compresses a video sequence into three layers: one base layer and two enhancement layers. The base layer is of low resolution and is encoded with the traditional prediction based codec, which can be compatible with existing video standards. Two fine-grain SNR scalable enhancement layers are encoded to improve the visual quality in different resolutions with PFGS technique [12]. Since a second motion compensation loop with high quality references is adopted in the PFGS technique, even though the base layer is low quality video at low resolution, the enhancement layers can still maintain good coding efficiency. The basic idea about the hybrid spatial and SNR scalable coding was first proposed in [13], but this paper develops a more flexible and efficient macroblock-based framework. In order to further improve coding efficiency and reduce drifting errors at the HR enhancement layer, the proposed framework defines five coding modes with different forms in motion compensation and reconstruction.

Furthermore, a mode decision algorithm is developed to select the suitable coding mode for each HR enhancement macroblock.

This paper is organized as follows. Section 2 discusses the proposed framework with hybrid spatial and fine-grain SNR scalabilities in details. Section 3 defines five coding modes for the HR enhancement layer coding. The distinctions among them are different references used for motion compensation and reconstruction. In order to select the suitable coding mode for each HR enhancement macroblock, the algorithm for mode decision is developed in section 4. Section 5 gives experimental results and comparisons between the proposed framework and the traditional spatial scalable video coding scheme. Finally, section 6 concludes this paper.

2. VIDEO CODING FRAMEWORK WITH HYBRID SPATIAL AND FINE-GRAIN SNR SCALABILITIES

A typical framework with hybrid spatial and fine-grain SNR scalabilities is illustrated in Fig. 2, where small boxes indicate LR video, and large boxes indicate HR video. The top row in Fig. 2 is the base layer, and each of other rows denotes one bit plane either at the LR enhancement layer or at the HR enhancement layer. Since both enhancement layers adopt the PFGS technique, each enhancement layer can be predicted from the current frame at the base layer and the previous frame at the same resolution enhancement layer. With bit plane coding technique, the residue between the original LR video and the LR prediction is encoded to form the LR enhancement bitstream, while the residue between the original HR video and the HR prediction is encoded to form the HR enhancement bitstream. Since the LR enhancement layer is encoded with the same technique proposed in [14], this paper focuses on the discussion on how to efficiently compress the HR enhancement layer.

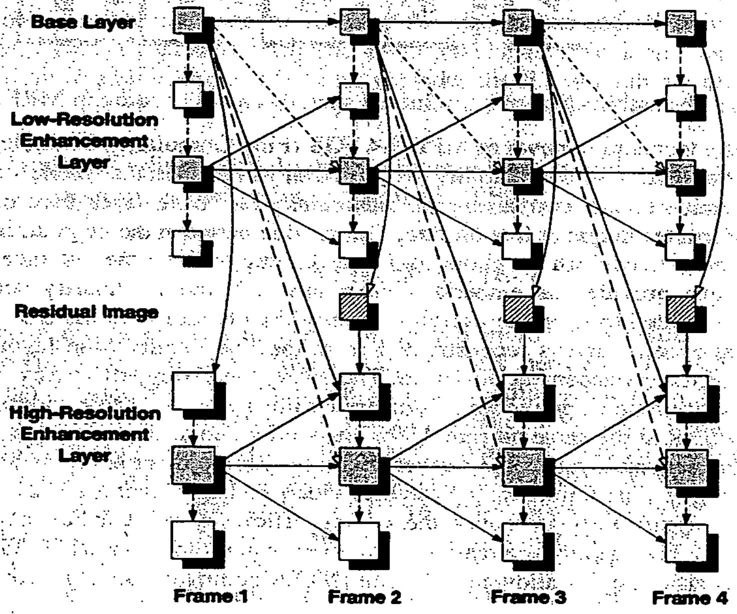

Base Layer
Low-Resolution Enhancement Layer
Residual Image
High-Resolution Enhancement Layer
Frame 1
Frame 2
Frame 3
Frame 4

Fig. 2: The proposed framework with hybrid spatial and fine-grain SNR scalabilities.

As illustrated in Fig. 2, the first frame of each GOP (Group of Picture) is encoded as forward prediction frame at the HR enhancement layer. The reference is derived from the current reconstructed base layer (at low resolution) by interpolation, where the bilinear method is used for simplicity. In order to obtain the fine granularity scalability, the bit plane coding technique is applied to compress the predicted DCT coefficients into an embedded bitstream instead of the traditional VLC coding. The rest of frames in this GOP are predicted with two references. One is the low quality reference constructed from the previous LR base layer after bilinear interpolation. Another is the high quality reference constructed from the HR enhancement layer. In fact, there are two methods for construction of the low quality reference, from the base layer or from the LR enhancement layer. If the low quality reference is constructed from the LR enhancement layer, the base layer video is first improved at low resolution and then switch to high resolution. This would decrease the coding efficiency at HR video, because it is more efficient to allocate the bits used for the LR enhancement layer coding directly to the HR enhancement layer coding. The comparisons between these two methods are given in the section 5. Thus, the proposed framework prefers to construct low quality reference from the base layer. Since the two enhancement layers at different resolutions share the same base layer, the proposed framework can readily switch from one enhancement layer to the other.

The proposed framework separates the reconstructed base layer of the current frame into two parts: previous temporal prediction and the residual image encoded in the current frame. The residual images as shown by the boxes with slant stripes in Fig. 2 are reconstructed from the DCT coefficients encoded in the base layer by dequantization and IDCT transform. Each of these two parts after bilinear interpolation is forwarded to the HR enhancement layer. The residual image is an option in the HR enhancement coding. When the HR enhancement layer is predicted from the low quality reference, the encoded residual image is always subtracted from the predicted error image. When the HR enhancement layer is predicted from the high quality reference, if the residual image can further decrease the energy of the predicted error image, the residual image is subtracted from the predicted error image. Otherwise the predicted error image is directly forwarded to the bit plane coding. This makes the proposed framework more flexible.

Solid lines and solid arrows in Fig. 2 denote temporal predictions between adjacent frames. In general, the same prediction should be used in both motion compensation and reconstruction. However, the high quality reference contains several bit planes from the HR enhancement layer, and if those bit planes cannot be correctly and completely transmitted to client under certain network conditions, this would inevitably cause drifting errors. In order to reduce the drifting errors, although some frames at the HR enhancement layer are predicted from the previous high quality reference, the current high quality references in these frames are still reconstructed from the previous low quality reference as shown by dash lines with hollow arrows in Fig. 2. This means different references for motion compensation and reconstruction. In the scalable video coding, because the bit rate of the base layer is very low, it is reasonable to assume that the base layer is always available at the decoder. Thus, this method can effectively reduce the drifting errors propagated from previous frames by utilizing the same base layer reference for reconstruction at both the encoder and the decoder. Dash lines with solid arrows in Fig. 2 denote the relation between bit planes.

Even though the base layer is of low resolution, the proposed framework can still maintain good coding efficiency at the HR enhancement layer due to the high quality reference. Moreover, since the bit plane coding and drifting reduction techniques are used at the enhancement layers, the enhancement layer bitstreams can be arbitrarily truncated to fit in the available channel bandwidth. Therefore, the proposed framework not only provides spatial scalability but also provides fine granularity SNR scalability.

Fig. 3 is the block diagram of the proposed encoder with hybrid spatial and fine-grain SNR scalabilities corresponding to the architecture illustrated in Fig. 2. This diagram omits the part for the LR enhancement layer coding for simplicity. At the base layer, input LR video is encoded by traditional motion compensation and DCT transform techniques, which can be compatible to existing video standards. By referencing the LR reference stored in the *Frame0* module, the motion estimation module first estimates the motion vectors for the LR video at the base layer. The predicted image p_b is calculated from the LR reference ref_b using the estimated motion vectors. After DCT and quantization, the quantized DCT coefficients are encoded to generate the base layer bitstream with traditional VLC coding. With dequantization and IDCT transform, the DCT coefficients encoded at the base layer are reconstructed as the residual image $\tilde{x}_b$. The reconstructed base layer is equal to the sum of the residual image $\tilde{x}_b$ and the prediction p_b, which is the reference for the next frame.

The residual image $\tilde{x}_b$ and the predicted image p_b at the base layer are forwarded to the HR enhancement layer. After bilinear interpolation, the corresponding residual image and the low quality reference are $\tilde{x}_{Hb}$ and p_{Hb}, respectively. There is another motion estimation module at the HR enhancement layer. The high-resolution motion vectors are obtained by referencing to the high quality reference ref_{He}. Not all high-resolution motion vectors are encoded in the HR enhancement layer bitstream. As we will see later, at macroblock level, some macroblocks can be predicted and reconstructed from the low quality reference p_{Hb}, and the corresponding motion vectors are the same as that at the base layer. There is no need to encode the high-resolution motion vectors in this case.

The high quality prediction p_{He} is obtained from the high quality reference using the high-resolution motion vectors. In other words, the proposed framework has two temporal predictions p_{Hb} and p_{He} at the HR enhancement layer. Combining with the residual image $\tilde{x}_{Hb}$, there are actually three methods for prediction at the HR enhancement layer: low quality prediction p_{Hb} plus $\tilde{x}_{Hb}$, high quality prediction p_{He} and high quality prediction p_{He} plus $\tilde{x}_{Hb}$. The switch S_1 in Fig. 3 selects the prediction for motion compensation, whereas the switch S_2 selects the prediction for reconstruction.

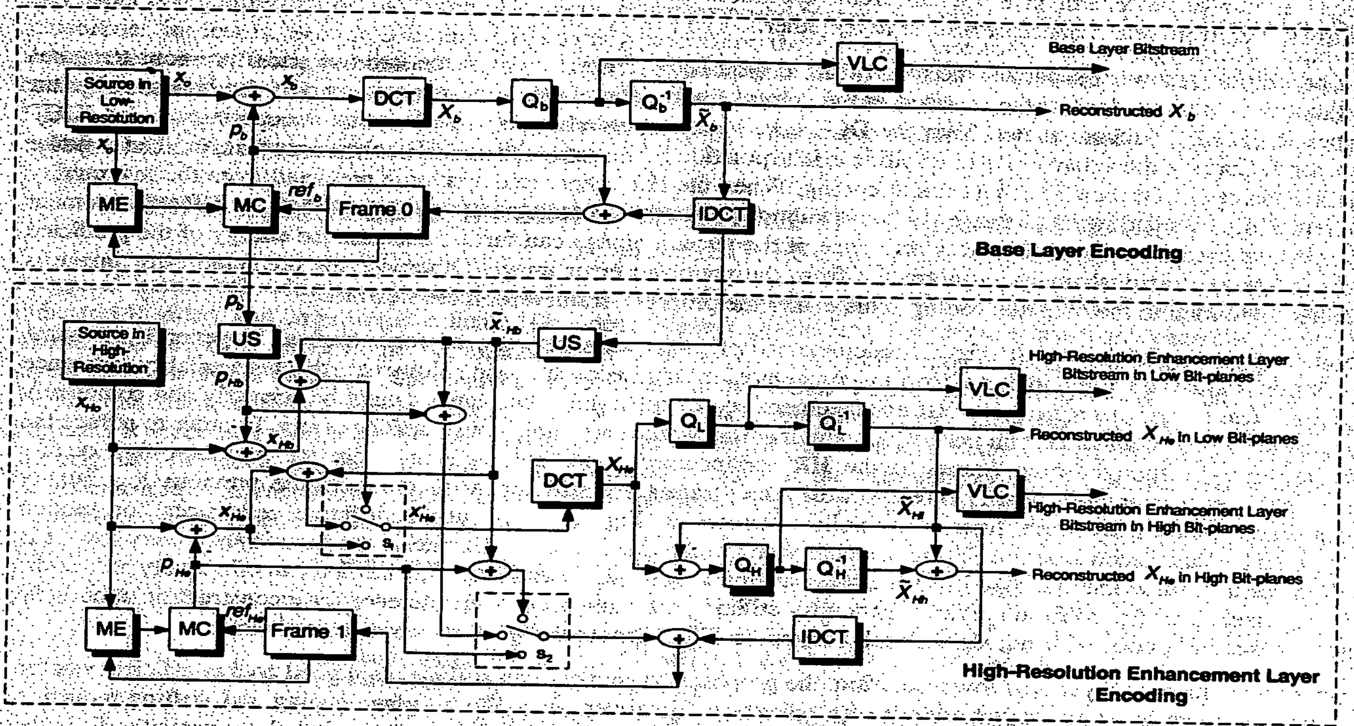


Fig. 3: The block diagram of the proposed encoder with hybrid spatial and fine-grain SNR scalabilities.

3. FIVE MODES FOR MACROBLOCK-BASED CODING AT THE HIGH-RESOLUTION ENHANCEMENT LAYER

As mentioned above, there are three methods for constructing the prediction for motion compensation and reconstruction at the HR enhancement layer. Moreover, to reduce the drifting errors, the prediction for motion compensation can be different from the prediction for reconstruction. The framework shown in Fig. 2 can be either operated at frame level or at macroblock level. As suggested in [14], since macroblock-based coding can provide both flexibility and much improved coding efficiency, we adopt the macroblock-based approach in this paper. In order to flexibly control motion compensation and reconstruction at macroblock level, this section defines five coding modes as shown in Fig. 4 for coding macroblocks at the HR enhancement layer. In other words, each macroblock at HR enhancement layer can select the individual method for motion compensation and reconstruction. The coding mode information of each macroblock is encoded in the frame header by simple VLC coding.

In Mode 1, the HR image is compensated and reconstructed both from the low quality prediction p_{Hb} plus the residual image $\tilde{x}_{Hb}$. Since the bit rate of the base layer is very low, it is reasonable to assume that the base layer bitstream can always be correctly transmitted to client with appropriate error protection. In other words, the low quality prediction p_{Hb} and the residual image $\tilde{x}_{Hb}$ are always available at the decoder. Therefore, this mode does not cause any drifting errors. However, since the prediction is derived from the LR base layer, the coding efficiency of this mode is usually low.

In Mode 2, the HR image is compensated and reconstructed both from the HR prediction p_{He} plus the residual image $\tilde{x}_{Hb}$. Since the high quality prediction is used at the HR enhancement layer, if all macroblocks are encoded with this mode, the proposed framework can achieve high coding efficiency as long as the high quality prediction is available at the decoder. However, since the high quality prediction contains several bit planes from the HR enhancement layer, they may be dropped when channel conditions somehow deteriorate. In this case, the decoder has to use the corrupted high quality prediction or the low quality prediction instead. This would inevitably cause the drifting errors. As a result, the decoded visual quality would be rapidly deteriorated across frames until another key frame.

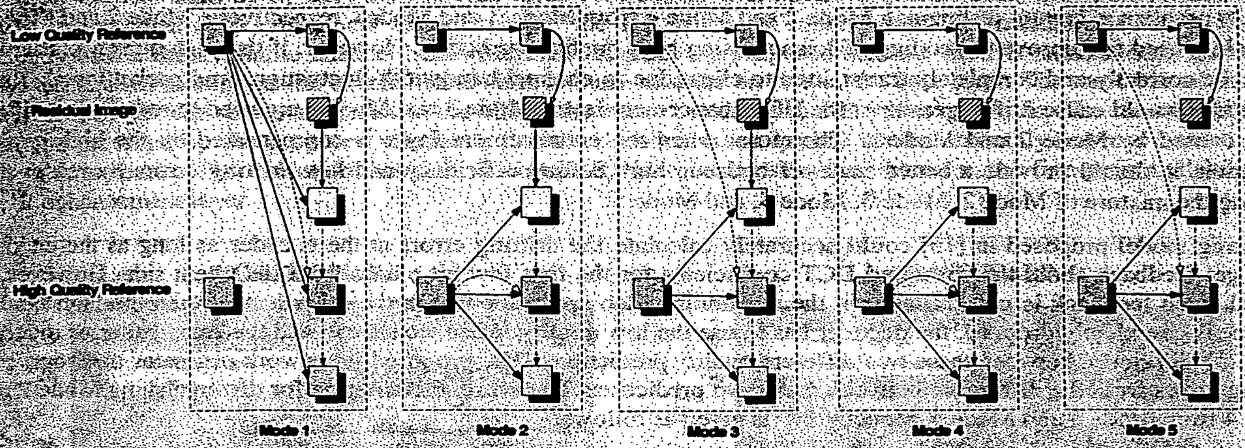


Fig. 4: The coding modes for the HR enhancement macroblock coding.

Mode 3 can effectively reduce the drifting errors at the HR enhancement layer. In this mode, the HR image is compensated from the high quality prediction p_{He} plus the residual image $\tilde{x}_{Hb}$, but reconstructed from the low quality prediction p_{Hb} plus the residual image $\tilde{x}_{Hb}$. When a macroblock is encoded with this mode, since the low quality prediction is always available at the decoder, both the encoder and the decoder can always use the same prediction for reconstruction. Therefore, the drifting errors propagated from the previous frames caused by the high quality prediction can be terminated by this method. This method also has a negative effect on the coding efficiency of the HR enhancement layer. The current reconstructed high quality reference for the next frame cannot obtain the best quality it could get because the mismatch between motion compensation and reconstruction. Fortunately, this kind of loss in coding efficiency is generally small since shorter bitstream compensate part of the loss. Moreover, we can also reconstruct a better quality macroblock just for display purpose if the high quality prediction is also available at the decoder at high bit rates.

In the first three coding modes mentioned above, the residual image $\tilde{x}_{Hb}$, encoded at the base layer is always involved in the HR enhancement layer coding. However, when the HR enhancement layer is compensated from the high quality prediction, sometimes the residual image $\tilde{x}_{Hb}$ encoded at the base layer may increase the predicted errors due to different resolutions and different motion vectors at the base layer and the HR enhancement layer. Therefore, Mode 4 and Mode 5 are proposed in the HR enhancement layer coding. Compared with Mode 2 and Mode 3, the only difference is that now the residual image $\tilde{x}_{Hb}$ is not subtracted from the predicted error image.

4. MODE DECISION AT THE HIGH-RESOLUTION ENHANCEMENT LAYER

The purpose of mode decision is to determine the coding mode for each macroblock at the HR enhancement layer. The proposed framework is expected to maintain high coding efficiency at high bit rates and reduce the drifting errors as much as possible at low bit rates by selecting appropriate coding mode for each macroblock. As discussed above, there are three types of prediction at the HR enhancement layer. Thus, the criterion for selecting prediction is given as follows,

$$\min\left(\left\|x_{Ho}-p_{Hb}-\tilde{x}_{Hb}\right\|,\ \left\|x_{Ho}-p_{He}-\tilde{x}_{Hb}\right\|,\ \left\|x_{Ho}-p_{He}\right\|\right) \qquad (1)$$

Here x_{Ho} is the original HR video. When the HR enhancement layer is compensated from the low quality prediction plus the residual image, the predicted error image encoded is the first item in criterion (1). When it is compensated from the high quality prediction plus the residual image, the predicted error image encoded is the second item. When it is only compensated from the high quality prediction, the predicted error image encoded is the third item. Just like in motion estimation, this criterion utilizes the sum of absolute difference (SAD) to measure the efficiency of prediction. If the first item of criterion (1) is the least for a certain macroblock at the HR enhancement layer, this macroblock is encoded with Mode 1. If the second item is the least, the corresponding macroblock is encoded with Mode 2 or Mode 3. If the third item of the criterion is the least, the corresponding macroblock is encoded with Mode 4 or Mode 5.

The key problem is how to distinguish Mode 2 and Mode 4 from Mode 3 and Mode5, respectively. Mode 2 and Mode 4 are mainly used to improve the coding efficiency using high quality prediction. However, if the high quality prediction cannot be correctly and completely transmitted to client due to channel bandwidth fluctuations and packet losses, these two modes would cause drifting errors at the HR enhancement layer. Mode 3 and Mode 5 are used to reduce the drifting errors caused by Mode 2 and Mode 4. Therefore, when the enhancement layer is compensated by the high quality prediction, it should provide a better trade-off between high coding efficiency and low drifting errors by reasonably utilizing the mixture of Mode 2, Mode 3, Mode 4, and Mode 5.

A drifting model proposed in [15] could accurately calculate the drifting errors at the encoder as long as the encoder could get feedback about the corrupted DCT coefficients for the first few bit planes from the client side. However, in general, such feedback is not available in the streaming video applications. On the other hand, when the high quality referece is completely dropped, the decoder has to use the low quality reference instead and the maximal error incurred is just the difference between these two different quality predictions. The proposed framework estimates the worst case drifting errors from the difference between these two predictions.Therefore, the criterion for distinguishing Mode 2 and Mode 4 from Mode 3 and Mode 5, repectively, is given as follows,

$$\left\| p_{He} - p_{Hb} \right\| > k \times \left\| x_{H0} - p_{He} \right\| \qquad (2)$$

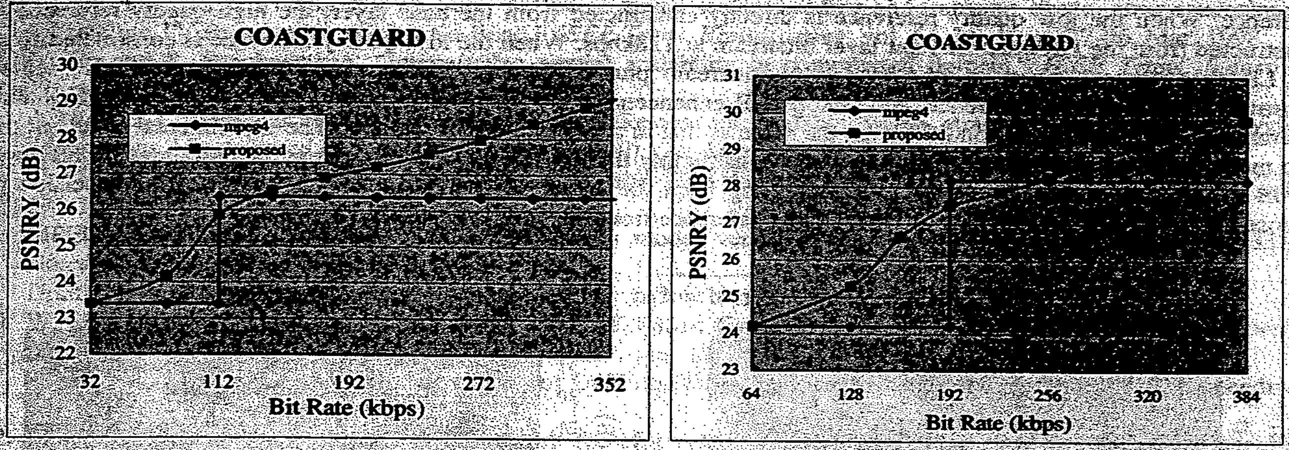
The left side of inequility (2) is the drifting errors caused by replacing the high quality prediction with the low quality prediction. The right side of inequility (2) is the given threshold for the allowable quality losses of the HR enhancement layer at lower bit rates. The factor k is an adjustable parameter to provide the balance between drifting errors and the coding efficiency. In other words, if the drifting error at the current macroblock is larger than the allowable quality losses, this macroblock is encoded with Mode 3 or Mode 5 for drifting reduction; otherwise this macroblock is encoded with Mode 2 and Mode 4 for high coding efficiency. Therefore, the proposed framework can easily select the macroblock coding mode by combining criteria (1) and (2).

5. EXPERIMANTAL RESULTS

This section verifies the proposed framework with hybrid spatial and fine-grain SNR scalabilities through extensive experiments. The performance of the proposed framework (denoted as the *proposed* scheme) is compared with that of the MPEG-4 traditional spatial scalable coding scheme (denoted as the *traditional* scheme). The MPEG-4 test sequences Forman and Coastguard are used in the experiments. The base layer is encoded with the MPEG-4 advanced simple profile (ASP) at two different bit rates, 32kbps and 64kbps. TM5 rate control is used in the base layer encoding. For each sequence, only the first frame is encoded as I frame, and the rest of frames are encoded as P frames. The encoding frame rate is 10Hz. The LR video is in QCIF format, and the HR video is in CIF format. The ranges of motion vectors are set as ±15.5 pixel for the low-resolution video and ±31.5 pixel for the high-resolution video. There is no rate control at both the LR enhancement layer and the HR enhancement layer in the proposed scheme, and they can be truncated at any point according to the available current bandwidth. Since LR enhancement layer is essentially the same as in [14], we will not discuss it here. We should still note that with the same base layer, when the users or devices desire a low resolution but higher quality video, the LR enhancement layer can be readily transmitted to the client to improve the video quality. In the following discussions, we will focus on the smooth transition from LR video to HR video using the HR enhancement layer.

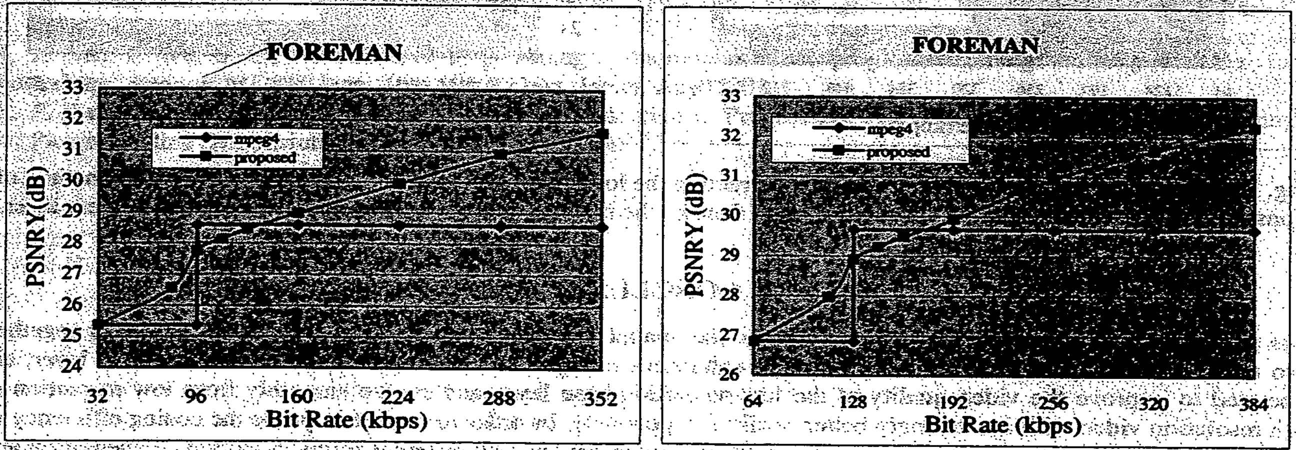
The experimental results of the Coastguard sequence are given in Fig. 5. In experiment (a), the bit rate of the base layer is 32kbps. The bit rate of the enhancement layer is 80kbps in the traditional scheme. In other words, SBR (Switch Bit Rate) is 112kbps in experiment (a). In experiment (b), the bit rate of the base layer is 64kbps, and the bit rate of the enhancement layer in the traditional scheme is 128kbps. SBR is 192kbps in experiment (b). In these two experiments, the HR enhancement layer bitstream of the proposed scheme can be arbitrarily truncated according to the available channel bandwidth. For example, when the available channel bandwidth is 128kbps, if the bit rate of the base layer is 64kbps, the HR enhancement layer bitstream is truncated to 64kbps. The results of the Foreman sequence are given in Fig. 6. The base layer bit rate in experiment (a) is 32Kbps, and the enhancement layer bit rate in traditional scheme is 64Kbps. The base layer bit rate in (b) is 64Kbps, and the enhancement layer bit rate in traditional scheme is 64Kbps. As shown in Fig. 5 and Fig. 6, the traditional scheme only provides two choices: either low resolution or high resolution.

Each layer in the traditional scheme is optimized at a fixed bit rate. Once encoded, the enhancement layer bitstream has to be either completely transmitted or completely dropped. When channel bandwidth is more than the bit rate of the base layer but less than SBR, only the LR base layer bitstream is transmitted to client. Until the channel bandwidth increases more than SBR, the client cannot receive the HR video. In other words, the traditional scheme cannot efficiently utilize the available channel bandwidth. However, in the proposed scheme, the HR enhancement bitstream can be transmitted as soon as the channel bandwidth is more than the bit rate of the base layer. The more bits of the HR enhancement layer the server transmits to the client, the better the decoded HR video quality is. The most significant advantage provided by the proposed scheme is that the decoded video quality can be smoothly improved as channel bandwidth increases. The evolution from LR video to HR video is also seamless thanks to the fine-grain scalability in the HR enhancement layer.



(a) The bit rate of the base layer is 32kbps. (b) The bit rate of the base layer is 64kbps

Fig. 5: Comparisons between the traditional spatial scalable coding (MPEG4) scheme and the proposed scheme.



(a) The bit rate of the base layer is 32kbps. (b) The bit rate of the base layer is 64kbps.

Fig. 6: Comparisons between the traditional spatial scalable coding (MPEG4) scheme and the proposed scheme.

At the base layer bit rate, the proposed scheme and the traditional scheme have the same decoded visual quality. As the channel bandwidth increases from the base layer bit rate to SBR, the proposed framework can send more bits at the HR enhancement layer to improve the decoded visual quality. But the traditional scheme still has to send the base layer bitstream only. Therefore, the decoded quality of the proposed scheme is significantly better than that of the traditional scheme. When the channel bandwidth is equal to SBR, the traditional decoder starts to decode the enhancement layer bitstream. The decoded quality of the traditional scheme is about 0.7dB higher at SBR than that of the proposed scheme. The coding efficiency loss in the proposed scheme is exchanged for the fine-grain scalable property at the HR enhancement layer. Because of channel bandwidth fluctuations, the traditional scheme cannot immediately switch to the enhancement bitstream when the channel bandwidth reaches exactly the SBR. Therefore, the little coding efficiency loss at SBR can be ignored in most practical applications. As the network bandwidth continues to increase, the proposed scheme can further improve the decoded visual quality with more bits at the HR enhancement layer transmitted to client. On the other hand, the maximal bit rate provided by the traditional scheme is SBR even if there may be more channel bandwidth available. The decoded video quality of the proposed scheme is higher than that of the traditional scheme again at high bit rates.

The following experiments compare the two different methods for constructing the low quality reference as discussed in section 2. The bit rate of the base layer is 64kbps, and an additional 128kbps is used to construct the high quality reference. When the low quality reference is directly constructed from the base layer, the high quality reference is constructed from the HR enhancement layer bitstream at 128kbps. When the low quality reference is constructed from the LR enhancement layer, 64kbps base layer bitstream plus 32kbps LR enhancement layer bitstream is used to construct the low quality reference, and 96kbps HR enhancement layer bitstream is used to construct the high quality reference. Other experiment conditions are the same as that in the previous experiments. The experimental results of the two methods are given in Fig. 7. The scheme of constructing the low quality reference from the base layer brings up to 1.0dB PSNR gain over the scheme of constructing from the LR enhancement layer. This validates our decision to choose the method of constructing the low quality reference from the base layer section 2.

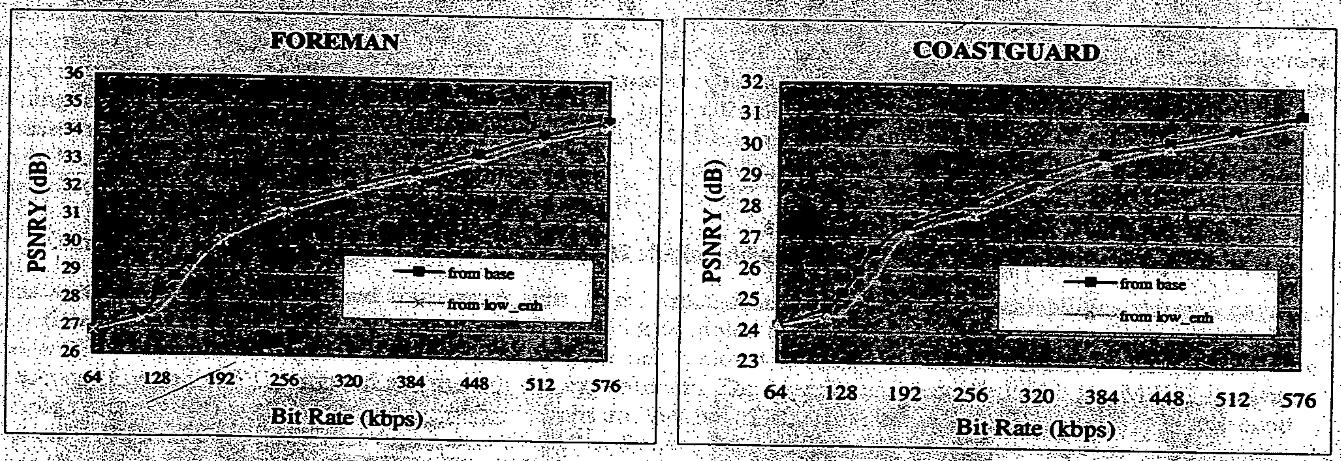


Fig. 7: Comparison between two methods of constructing the low quality reference. "from base" denotes the method of constructing from the base layer, "from low_enh" denotes the method of constructing from the LR enhancement layer.

6. CONCLUSIONS

This paper presents a novel framework to combine the spatial scalability and the fine-grain SNR scalability together. Two enhancement layers, i.e., the low-resolution enhancement layer and the high-resolution enhancement layer, are generated to improve the video quality of the low-resolution base layer and evolve smoothly from low resolution to high resolution video with increasingly better quality, respectively. In order to further improve the coding efficiency of the high-resolution enhancement layer, five macroblock coding modes with different forms in motion compensation and reconstruction are proposed in this paper. Furthermore, a mode decision algorithm is developed to select the appropriate coding mode for each high-resolution enhancement macroblock. Compared with the traditional spatial scalable coding scheme, the proposed hybrid spatial and fine-grain SNR scalable coding scheme not only provides the spatial scalability but also provides the fine granularity quality scalability at the same resolution. This is an important step towards the so called universal scalable video coding where spatial, temporal and fine granularity SNR scalabilities are integrated together to provide ubiquitous video access on any devices through any networks. With the new functionality, the quality loss in the proposed scheme is small at the SBR and significant quality improvement is obtained when the channel bandwidth increases.

REFERENCES

1. B. J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374-1387, Dec. 2000.
2. U. Benzler, "Spatial scalable video coding using a combined subband-DCT approach," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1080-1087, Oct. 2000.
3. M. Domanski, A. Luczak, and S. Mackowiak, "Spatio-temporal scalability for MPEG video coding," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1088-1093, Oct. 2000.
4. Y. Hsu, Y. Chen, C. Huang, and M. Sun, "MPEG-2 spatial scalable coding and transport stream error concealment for satellite TV Broadcasting Using Ka-Band," *IEEE Trans. Broadcasting*, vol. 44, no. 1, pp. 77-86, Mar. 1998.
5. Q. Hu and S. Panchanathan, "Image/video spatial scalability in compressed domain," *IEEE Trans. Industrial Electronics*, vol. 45, issue. 1, pp. 23-31, Feb. 1998.
6. A. Lallet, C. Dolbear, J. Hughes, P. Hobson, "Review of scalable video strategies for distributed video applications," *Distributed Imaging (Ref. No. 1999/109), IEE European Workshop*, vol. 2, pp. 1-7, 1999.
7. T. Chiang, H. Sun, and J. W. Zdepski, "Spatial Scalable HDTV coding," *Proc. IEEE, Image Processing*, vol. 2, pp. 571-574, 1995.
8. MPEG Video Group, "MPEG-4 video verification model version 16.0," ISO/IEC JTC1/SC29/WG11 N3312, Norrdwijkerhout, Netherlands, Mar. 2000.
9. Draft ITU-T Recommendation H.263, "Video coding for low bit rate communication," Jan. 1998.
10. W. Li, "Streaming video profile in MPEG-4," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on Streaming Video*, vol. 11, no. 3, pp. 301-317, Mar. 2001.
11. F. Ling, W. Li and H. Sun, "Bitplane coding of DCT coefficients for image and video compression", Proceedings of SPIE VCIP'99, San Jose, Jan.25-27, 1999.
12. F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on Streaming Video*, vol. 11, no. 3, pp. 332-344, Mar. 2001.
13. Q. Wang, F. Wu, S. Li, Y. Zhong, and Y.-Q. Zhang, "Fine-granularity spatially scalable video coding", *IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1801-1804, Salt Lake City, May, 2001.
14. X. Sun, F. Wu, S. Li, W. Gao, and Y.-Q. Zhang, "Macroblock-based progressive fine granularity scalable video coding," *Proc. ICME 2001*, pp. 461-464, Tokyo, Japan, Aug. 2001.
15. F. Wu, S. Li, B. Zeng, Y.-Q. Zhang, "Drifting Reduction in Progressive Fine Granular Scalable Video Coding", Picture Coding Symposium (PCS), Seoul, April 2001.